

## FERMILAB-Conf-94/084

# Sloan Digital Sky Survey

S.M. Kent, C. Stoughton, H. Newberg, J. Loveday, D. Petravick, V. Gurbani, E. Berman and G. Sergey

> Fermi National Accelerator Laboratory P.O. Box 500, Batavia, Illinois 60510

> > R. Lupton

Princeton University Ivy Lane, Princeton, New Jersey 08544

## April 1994

To appear in the Proceedings of the 3rd Annual Conference on Astronomical Data Analysis Software and Systems, Victoria, British Colombia, October 13-15, 1993

## Disclaimer

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

## Sloan Digital Sky Survey

Stephen M. Kent, Chris Stoughton, Heidi Newberg, Jonathan Loveday, Don Petravick, Vijay Gurbani, Eileen Berman, Gary Sergey

Experimental Astrophysics Group, Fermi National Accelerator Laboratory, MS 127, PO Box 500 Batavia, IL 60510 USA

Robert Lupton

Princeton University, Peyton Hall, Ivy Lane, Princeton, NJ 08544 USA

Abstract. The Sloan Digital Sky Survey will produce a detailed digital photometric map of half the northern sky to about 23 magnitude using a special purpose wide field 2.5 meter telescope. From this map we will select  $\sim 10^6$  galaxies and  $10^5$  quasars, and obtain high resolution spectra using the same telescope. The imaging catalog will contain  $10^8$  galaxies, a similar number of stars, and  $10^6$  quasar candidates.

#### 1. Introduction

The partners in the SDSS are the University of Chicago, Princeton University, the Institute for Advanced Study, Fermilab, Johns Hopkins University, and the Japanese Promotion Group. The specific goals of the survey are as follows:

- 1. In  $\pi$  steradians of the North Galactic Pole: (a) obtain a photometric survey in 4 or 5 filters to R=23 (5 $\sigma$ ); (b) obtain redshifts for all galaxies to B=19; (c) obtain redshifts for all quasars to B=20.
- 2. In a strip  $2^{\circ} \times 50^{\circ}$  of the South Galactic Pole: (a) obtain a deep photometric survey to R=25; (b) obtain redshifts for all galaxies to B=20; (c) obtain redshifts for all guasars to B=21.
- 3. If feasible, obtain a best effort imaging survey of the Galactic Plane.

The survey will use a new f/5 2.5 meter telescope of altitude-azimuth design that is under construction at Apache Point Observatory (APO) in New Mexico. The telescope is a modified Ritchey-Cretien design that uses two corrector lenses near the focal plane to achieve a 3° field of view with no distortion.

Imaging will be done using a camera that consists of a mosaic of 52 CCDs. Thirty of these CCDs are Tektronix  $2048 \times 2048$  arrays that are used for the primary imaging observations, arranged in an array of 6 columns with 5 CCDs per column (Figure 1). Each CCD in a column has a different filter with the following wavelengths: u: 3506; g: 4734; r: 6270; i: 7691; z: 9247.

The imaging survey will be conducted in drift scan mode. The telescope will be actively tracked so that a given piece of the sky trails along the 5 CCDs

of a column in succession. The transit time of a single CCD will be 55 seconds and the time to cross the array will be about 7 minutes. The columns of CCDs are spaced by slightly less that one CCD width; thus, 2 successive interlaced scans of the telescope will produce a completely filled image of a strip of the sky 2.5 degrees wide.

A total of 22 small CCDs, leading and trailing the main imaging CCDs, provide the astrometric calibration. These CCDs will tie bright (V < 9) stars with known astrometric positions to fainter (V = 14) secondary stars. The desired accuracy is 0.2'' rms in each coordinate.

A separate 0.61 meter monitor telescope (MT) will be used for the photometric calibration. This monitor telescope will have a single CCD camera and a filter wheel box. The functions of this telescope are fourfold. First, it will set up a set of standard stars for the photometric calibration (the SDSS filters are not on any standard system). Second, during imaging observations, it will repeatedly observe the standard stars to monitor the atmospheric transparency. Third, it will observe a large number of patches in common with the 2.5 meter telescope, to calibrate the main imaging survey. Fourth, it will observe spectrophotometric standards to calibrate spectra.

Galaxy, quasar, and star targets will be selected from the imaging data for followup spectroscopy. The spectroscopy will be done with two multifiber spectrographs, each with a blue and red channel. The two spectrographs combined can measure 600+ objects simultaneously in the 3° field. The fibers will be positioned using drilled plates. The spectroscopic resolution is 3 Å, allowing velocity dispersions to be measured for the brighter galaxies. The exposure times will be on the order of 1 hour.

The galaxy survey is intended to be as complete as possible. Galaxies will be skipped only if they are so close so as to cause interference between fibers. Since the distribution of galaxies on the sky is highly variable, the plate centers will not be placed on a uniform grid but rather will be adjusted to increase overlap in regions of high target density.

The results from the survey will be distributed in the most convenient form available. The products include: tables of all objects found in the survey ( $\sim 200\times 10^6$ ) with parameters; postage stamps of all objects; tables of redshifts for all objects with spectra; reduced 1 dimensional spectra; and the 2 dimensional images from which spectra were extracted.

Writing and running the code to build this archive requires coordinated effort of approximately 10 computer professionals at Fermilab and 20 or so scientists at the six institutions. In the following sections, we discuss the organization of these software products, as well as the tools we have developed and are using to work together.

## 2. Online Systems

The online systems are the hardware and software at APO that collect and record data from the instruments. Figure 2 shows the major pieces of the online systems.

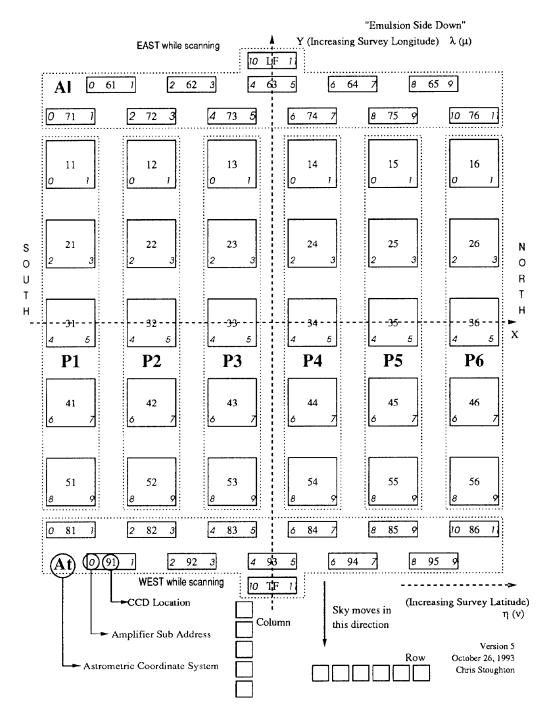


Figure 1. Focal plane layout for the CCD imaging camera

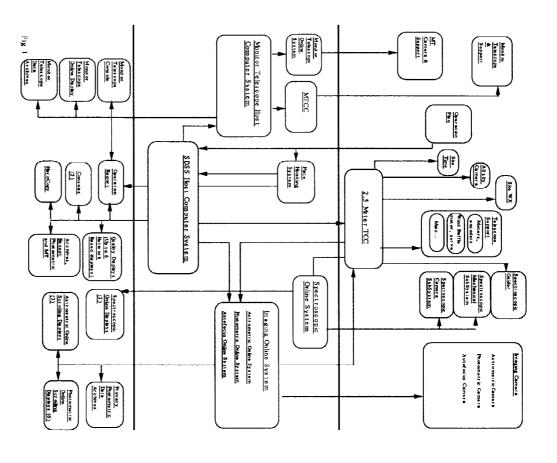


Figure 2. Online systems at Apache Point Observatory

#### 2.1. Imaging

The imaging camera produces the largest amount of data at the highest rate of all the instruments. MacKinnon (1993) describes the development of this system in some detail. We concentrate here on the organization of the data stream, online monitoring, and the data products written to tape.

The data acquisition and online processing are performed in MVME167 processor boards embedded in the data acquisition system. In addition to controlling the data flow between the CCD electronics, disk buffer, and tape drives, these processors will do a basic analysis of each frame. First, a histogram of all pixel values for each column of pixels is continuously computed. Periodically, the statistics of each histogram (the 25%, 50%, and 75% quartiles) are reported. This allows us to track the sky level. Second, bright objects are detected and copied out of the frames, allowing us to monitor the point spread function, object counts, and telescope tracking. Data are also sent to scrolling display monitors.

Before going to tape, the data are written to a hard disk buffer. We delay writing data from the leading chip in the column to tape until the corresponding position in the sky has scanned past all of the CCDs in the column. The five frames of different color are then recorded sequentially on tape. This disk buffer also allows us to selective copy frames to a UNIX environment for more detailed analysis and viewing. Although we do not support the full bandwidth in this mode, the flexibility enables us to track down problems as they arise.

At sidereal rate the CCDs are clocked at 37.5 rows per second. Including the overscan region, the data rate is 168.6 kbyte/sec for each CCD. The array of photometric chips (6 columns of 5 CCDs) produces data at just over 5 Mbyte/sec. To record at this rate we write data in parallel streams. The data from each column of 5 CCDs are written to two Exabyte tape drives to make primary and backup copies of the data simultaneously. If a single drive fails, we continue the scan and make the backup copy later. A total of 12 drives will record two copies of the photometric imaging data. The data products written by the imaging system are:

photometric frames: the pixel values from all of the photometric CCDs. postage stamps: the pixel values just around bright objects, extracted and measured online. This list is not guaranteed to be complete.

quartile arrays: the summary of the pixel intensity in each column, extracted from the histograms each frame.

imaging report: a summary of the imaging observing for the run – starting and stopping time, and summaries of the data quality (PSF profile, sky values) to give an overview of the observing conditions for the run.

instrument report contains data about special maintenance, such as replacing a CCD in the camera, that need to be tracked.

#### 2.2. Monitor Telescope

The camera electronics are a high speed version of those for the main imaging camera. This system is distinguished by its automated observing mode.

We will write the frames to disk during the night, and copy them (twice) to tape at the end of the night. The data products written by the monitor telescope system are:

monitor telescope frames are the frames for all the exposures.

monitor telescope report is the summary of the monitor telescope observations during the night, including which fields were completed, and also a summary of the atmospheric extinction measured throughout the night.

#### 2.3. Spectroscopic

The spectroscopic system controls the spectrograph, and acquires and monitors the spectra and associated calibration frames. We will extract and calibrate at least a subset of the spectra during the night, using IRAF.

The guiding system will be tightly coupled to the telescope control computer. It will read out small CCDs illuminated by coherent bundles of fibers placed on guide stars. The overall performance of this system, along with the measured flux and image shape, will also be recorded with the spectroscopic data. The data products written by the spectroscopic system are:

spectroscopic frames are the two-dimensional spectra (two or three exposures), along with any flats and calibration arcs that are routinely taken.

plugging report gives the correspondence from fiber number to hole number, to match spectra with objects.

spectroscopic report is a summary of the spectroscopic observations for the night, including the exposure time for each plate, as well as summary information from the guiding system.

## 3. Data Processing Pipelines

Data recorded on the mountain will all be shipped to Fermilab for subsequent processing. Figure 3 gives the overall organization of the offline data processing. We organize the processing into four pipelines: photometric, astrometric, monitor telescope, and spectroscopic. Our goal is to have the data processed "automatically" through the pipelines. The results are checked before committing them to the database. This check is then used to guide future observing plans to keep the survey uniform.

## 3.1. Astrometric Pipeline

The astrometric pipeline reads the postage stamp files from the astrometric and photometric CCDs. At the beginning of the survey we will construct a "great circle star catalog" of known astrometric standards (from the Hipparcos catalog, for example) converted to our survey coordinate system. Standards from this catalog are matched with stars measured in the astrometric CCDs to determine a primary calibration. Fainter stars on the astrometric CCDs transfer this calibration to the photometric CCDs. Because the final positions of fainter stars will not be available until after the photometric pipeline has run, provisions are made to recalibrate the output of the photometric pipeline at any time.

#### 3.2. Monitor Telescope Pipeline

The MT frames pipeline processes the two dimensional frames and produces lists of detected stars. The major steps are: flat field and bias correction, find the sky, locate the stellar images, measure instrumental aperture magnitudes, align frames in different colors, and (for the primary standards) identify the

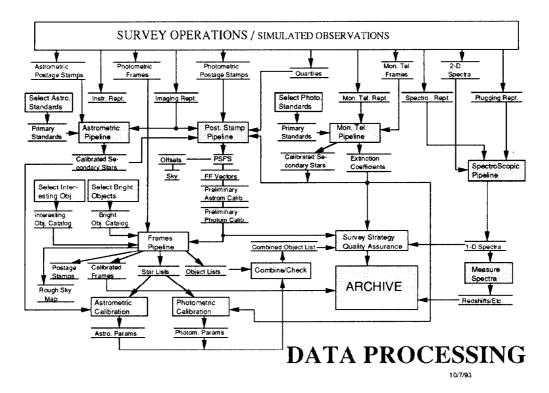


Figure 3. Overall design of the offline data processing

field. The output is a list of stars with instrumental magnitudes and colors. Then, the MT calibration pipeline runs on the primary standards to compute the extinction and instrumental coefficients.

### 3.3. Photometric Pipeline

The basic job of the photometric pipeline is to take the imaging data from the photometric CCDs and produce a list of objects. The "postage stamp pipeline" reads the postage stamps of bright stars generated by the imaging online system and determines: sky value, flat field, PSF, preliminary astrometric and photometric parameters. The "frames pipeline" reads all of the photometric frames to find and measure objects in each frame.

The processing steps are: correct frames; reject bad frames; find, measure, merge, bright objects; identify known objects; find objects; merge colors; and measure objects.

The output of the photometric pipeline is a list of the objects, their measured parameters, and the postage stamps (now a complete set of all detected objects) for each frame. The lists from all frames in a scan are combined. Final astrometric and photometric calibration parameters are calulated in a separate step once the astrometric and monitor telescope pipeline results are found satisfactory.

## 3.4. Spectroscopic Pipelines

The spectroscopic pipeline takes the data and calibration exposures and, along with the spectrophotometric standard measurements from the monitor telescope, extracts and measures the spectra.

We extract with following steps, using IRAF routines with minor modification to handle our data format: bias and throughput correction; aperture tracing; spectra extraction; dispersion correction; sky spectra extraction; sky subtraction; and median averaging.

These calibrated 1-d spectra are measured by: combining red and blue spectra; masking night sky lines; finding emission and absorption lines; identification with cross correlation.

#### 4. Software Tools

To coordinate the efforts of over two dozen people, computer professionals and scientists, at 6 institutions we define standards for the software development environment, and then support products for building the framework. A guiding principle is to make all the applications look as similar as possible.

#### 4.1. Standards

The standards for the hardware and software environment help us build robust, portable, documented code that is readily available to all members of the collaboration.

Platforms: We attempt to make our software reasonably portable. All code (with the exception of some online software) must run on both IRIX and SunOS operating systems. These are the operating systems in use at our member institutions.

Compilers: Virtually all new code is written in either ANSI C or C++. Some legacy code is written in Fortran. Code common to many applications is required to compile in 5 different compilers. This not only ensures that the code is portable, but also catches many subtle programming errors.

Software Distribution: Software distribution is via three tools: UPS (Unix Product Support) and UPR (Unix Product Remote) from Fermilab, and RCVS (Remote Concurrent Versioning System) from Stanford. UPS is a facility to maintain and support multiple versions and flavors (IRIX or SunOS) of our software products. UPR is a menu-driven facility that allows remote users at other institutions to fetch copies of the current versions of software products. RCVS is a source code management system that allows several people to "check out" a module for work simultaneously. Although this sounds a bit frightening at first, our experience (and the experience of others) is that when people continue to check in their changes regularly, any conflicts are detected by the system and are easy to resolve. The big advantage is that we avoid a gridlock situation where a small number of people are working on a module, locking out everyone else.

Documentation: We use TEXand LATEX converted to Postscript for printed documents, and the hyper text markup language (HTML) with a browser (Xmosaic) for reading online documents. A simple perl script converts HTML to

LATEX, so we need to maintian only one set of documents. This documentation is distributed or updated whenever the product is distributed.

#### 4.2. Framework

While it might have been natural to use IRAF as the framework for new code development, we chose not to do so for a couple of reasons. First, there was little IRAF experience within the collaboration, particularly at Fermilab. Second, IRAF is not well suited to the photometric pipeline where efficient control of system resources is needed. We created our own framework, drawing upon public domain software in many places. Virtually all applications, from the data acquisition systems to the database interface routines, are developed within this framework.

TCL (Berkeley) TCL (Tool Command Language) is the backbone. TCL is a programmable command line interpreter, akin to cl in IRAF. The major advantages are that it is highly portable (it even runs on the MVME 167 microprocessors in the online systems) and that it explicitly provides a clean interface for adding new commands implemented as blocks of C code. Online systems and data reduction pipelines are written as TCL verbs and scripts.

Tk (Berkeley) Tk (Took Kit) is a graphical user interface built with TCL commands. Simple screens to run online systems and, for example, to query the database, are built without compiling.

pgplot (Cal Tech) is a plotting package which we integrated as a set of TCL verbs and c modules.

**FSAOImage** (SAO) is an extension of the venerable SAOImage display program, implemented as a set of TCL verbs.

Libfits (Johns Hopkins) is basic FITS input and output.

SHIVA: Written at Fermilab, it binds these pieces together and supports a variety of data structures, such as linked lists and image regions.

IRAF: We will use IRAF in one place, the spectroscopic pipeline, for extraction of 1-d spectra from 2-d frames.

#### 5. Databases and Archives

The data archives are critical both for the operation of the survey and for the analysis of the final catalogs. All of the persistent data will be kept in an object oriented data base (OODB), with the possible exception of the primary photometric data and the corrected photometric frames.

We are currently using a commercial OODB (Versant) for the archive. We chose this company after a small evaluation of other alternatives, but are working to keep Versant-specific code isolated to facilitate upgrades and to facilitate moving to another vendor if necessary in the future.

There are many features of Versant, and of OODB technology in general, that are attractive. First, the structure of the database allows a natural interface to a high level language, in this case C++. The objects and methods of the database are implemented as C++ classes. Versant currently supports data bases on multiple machines, which is important as we access from different institutions and APO. Schema evolution is also supported to some extent, allow-

ing us to extend the definition of an object class without having to completely rebuild the database.

As an example of how Versant performs, we stored the ACRS catalog of 250,000 stars in Versant. Creating the database is done with simple UNIX commands. A user program was written to read the ACRS catalog from a FITS file and write it to the new database. During compilation, schema are generated which are put in the database to define which kinds of objects it will accept. It took the program 15 minutes on a modest UNIX workstation to load the full database. Another user program executed a Versant command to build an index on right ascension in the database. A third user program allows a user to query the database for objects in a specific RA and DEC range and return the results in a file. The resulting access time for the astrometric standards from one scan of the imaging camera is acceptably short. We are now in the process of loading the entire Guide Star Catalogs into a database to test Versant's performance on very large datasets.

#### 6. Conclusion

These systems work because we have defined all of the software systems to be modular. Each of the steps in the pipelines, as well as and the parts of the data acquisition systems, are developed independently once the interfaces are defined. We can install better versions of algorithms as they become available, test the robustness of each module before integration, and share module between systems, because of this design.

The Sloan Digital Sky Survey has a rather difficult goal: to construct a large, calibrated, consisted set of catalogs. Only by working together with such a set of software tools will it be possible for us to succeed.

Acknowledgments. Work at Fermilab supported by the U.S. Department of Energy under contract No. DE-AC02-76CH03000. The SDSS is partially supported by the Alfred P. Sloan Foundation.

## References

Gunn, J. E. and Knapp, G. R. 1992, PASP, 43, 267

Kron, R. G. 1992, ESO Conference and Workshop Proceedings, 42, 635

Bryan MacKinnon, B. et al. 1993, Conference Record of the Eighth Conference on Real Time Computing Applications in Nuclear, Particle, and Plasma Physics (IEEE), 329